Acta Technica *62* (2017), No. 5A, 305–312

Simulation of health insurance risk assessment model based on big data analysis

Xu Siyun[2,3], Liu Xiaoxia[2]

Abstract. Since the Reform and Opening, health insurance business has gradually become the backbone for promoting the development of financial industry. Limited by the existing marketing models of insurance companies, there is a certain risk in the acceptance of health insurance. Therefore, a health insurance risk assessment model based on big data analysis is designed. Based on big data analysis, according to the generality, similarity and broad impact of health insurance risks, health insurance risk assessment is conducted on the under-60s with the improved Lee-Carter model. The least squares method is used to estimate the unknown parameters of the model, and model time series are given. Finally, the maximum disease prevalence probability of the people at different age groups at a specific time is finally obtained. Simulation results show that health insurance risk assessment model based on big data analysis has strong resistance to invasion, low false alarm rate, short computation time and excellent overall performance.

Key words. Big data analysis, health insurance, risk assessment, model, simulation.

1. Introduction

In recent years, because of the policy of Reform and Opening, Chinese economy is in a trend of rapid increase. Insurance industry is gradually emerging and plays an important role in national economy. The public are benefited quite extensively.At the same time, insurance companies have come to realize that the marketing of health insurance is not a big profit with small capital, and the slightest mistake will lose a lot of money and lead to poor capital turnover.

[1]Acknowledgment - This paper is supported by province-level graduate professional practice base of Hebei Finance University and Ministry of Education of the People's Republic of China under Grant No.14YJC630044 and Hebei Planning Office of Philosophy and Social Science under Grant No. 201710120915 and soft science project of Hebei Science and Technology Department under Grant No. 164576472 and Baoding Science and Technology Bureau under Grant No. 16ZF014.

[2]Workshop 1 - Financial Synergy Innovation of Science And Technology Center in Hebei Province, Baoding,Hebei 071051,China

[3]Workshop 2 - University of International Business and Economics,Beijing,100029,China

Big data analysis can dig into the hidden data, so looking at health insurance from the perspective of big data analysis has a strong impetus to its development [1]. Bigger health insurance market share has the most far-reaching impact on insurance companies, making health insurance risk assessment become a decisive factor affecting the development of insurance companies. By big data analysis technology for risk assessment and simulation of health insurance, the relationship between the probability of people's illness and health insurance risks is to be solved and the awareness of insurance companies in our country about the risks is raised so as to provide a relatively safe operating environment for the stable development of China's health insurance.

2. Health insurance risk overview

The insured people of health insurance are all over the country, including insured enterprises and insured individuals [2]. Health insurance risk comes from the insured people's high probability of illness, which means that the insurance companies cannot afford the premium because there are too many insured persons need high premiums such as major diseases. The best solution to this problem is to distribute the insured people who need high premiums equally among different insurance companies. However, there is widespread competition among insurance companies, so average distribution cannot be realized generally. The insurers can only weigh the willingness to make insurances for these people [3]. In addition, insured people who need high premiums often do not have any illness early in the insured period. The risk assessments can be conducted on such people based on the disease prevalence in recent years in our country.

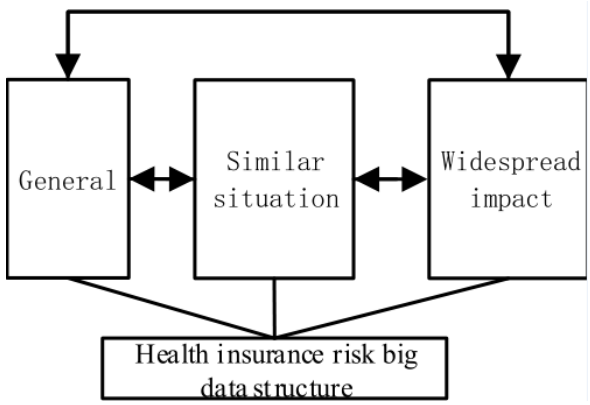

Fig. 1. Three major properties of health insurance risk

There are three major types of health insurance risks, as shown in Figure 1, which are general, similar situation and widespread impact [4], all of which are interrelated and influenced by each other and form the health insurance risk of complex big data structure. Generality refers to that health insurance risk is not caused by

the sickness of special people, but a general trend of social categories. Insurance companies cannot stop this trend and can only reduce their risk by choosing whether to accept the business or not. Similar situation means that the probability of illness among the public in each social stage is similar. As shown in Figure 2, during the 20 years from 1996 to 2016, the number of people suffering from major diseases in our country almost shows a linear upward trend.

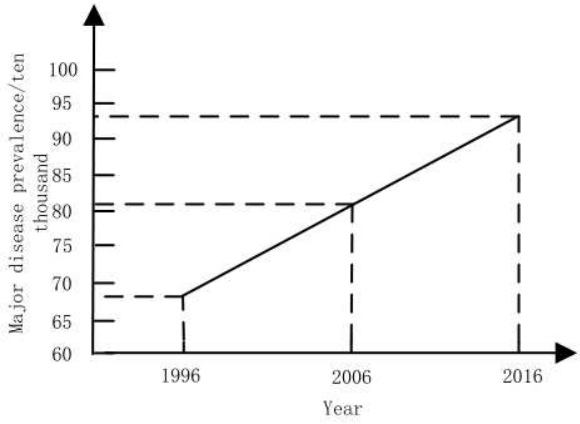

Fig. 2. Trend of the number of people suffering from major diseases in China in 1996-2016

Widespread impact refers to that different jurisdictions and different insurance companies are facing more severe illness. According to statistics by authoritative agencies, till the end of 2015, China's health insurance premium gap has reached as high as 100 billion Yuan [5].

3. Dynamic probability model based on big data analysis

In 1992, the United States used Lee-Carter model to assess the life expectancy of U.S. citizens and gave a predictive solution by means of matrix singular value decomposition method, which well described the dynamic change rule of population in the future. Therefore, in the design of health insurance risk assessment model based on big data, big data analysis of health insurance will be done through the improvement of Lee-Carter model. The dynamic model of diseased probability will be given and the parameters of diseased probability dynamic model will be analyzed item by item to realize the risk assessment of health insurance in China.

The expression of Lee-Carter model is:

$$\ln(m_{x,t}) = \alpha_x + \beta_x k_t + \varepsilon_{x,t} \tag{1}$$

In Lee-Carter model, m$x,t$is the death probability of people of x age at t time, αxrepresents the average death probability of people of x age, $\beta x$is the death probability of people at a current x age, and kt is time parameter, $\varepsilon x, t$ indicates the ran-

dom deviation. After the improvement, Lee-Carter model expression is completely adopted for the dynamic model of illness probability. mx, tmeans that people at x age have the probability of suffering from major diseases at t time, and αx is the average probability of illness prevailing in history and βx is the probability of people at the age of x suffering from major diseases. $kt$is the time parameter and $\varepsilon x, t$ indicates a random bias.

It can be seen that the illness probability dynamic model can well fit the tendency of the future illness probability and evaluate and predict the uncertainty of dynamic changes [6].

4. Health insurance risk assessment model simulation based on big data analysis

4.1. Assessment process design

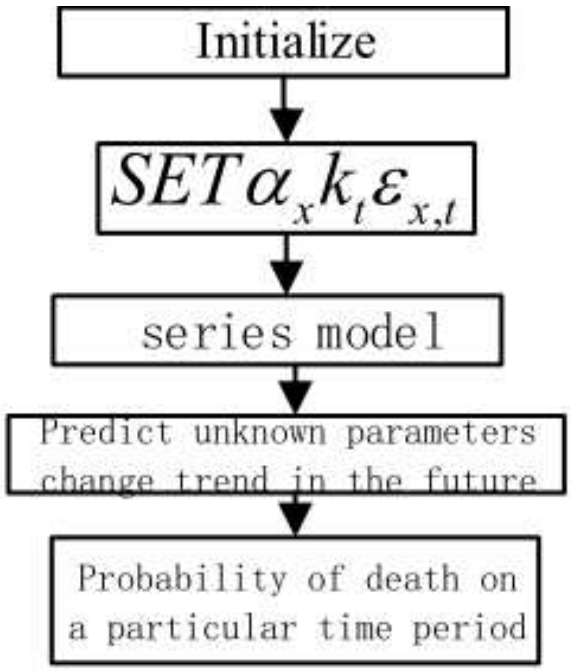

Fig. 3. Model assessment process

Figure 3 depicts the assessment process of health insurance risk assessment model based on big data analysis that involves risk assessment in three main steps. First, the unknown parameters in the model are estimated. The unknown parameters includeαx,kt and $\varepsilon x, t$[7]. Then model time series are given by analyzing the numerical value ofkt. The main function of model time series is to define the order of risk

assessment. The prediction of the change trend ofαx,kt and$\varepsilon x, t$is made according to the time series, and the probability of illness of the people at different ages at specific time is finally got. Insurance companies usemx, t to balance the acceptance of health insurance policyholders and to provide a relatively safe operating environment for its prosperity and stable development.

4.2. The unknown parameter estimation in the model

Based on the generality and similar situation of health insurance risks, there are several specific definitions of big data-based health insurance risk assessment models, including:

(1) Since health insurance risks have similar situational identities,$\alpha x$and$\beta x$are exactly equal to one another in the same annual interval. If the number of years covered more than one inmx, t, the value of$\alpha x$and$\beta x$are not necessarily equal;

(2) For random deviation$\varepsilon_{x,t}$, the arithmetic mean of random deviations for each annual assessment model is always equal to zero because of the generality of health insurance risks.

There are three unknown parameters in the health insurance risk assessment model based on big data and only one observed valuemx, t, which leads to multiple unknown parameter combination results if randomly estimateαx,$kt$and$\varepsilon x, t$, so that the implementing health insurance risk assessment needs to be conducted dialectically, and the forecast of future development trends has caused a lot of unnecessary troubles as well as reduced efficiency of risk assessment. Therefore, big data-based health insurance risk assessment model will use the least square method to estimate the specific values of the three unknown parameters, and give some model parameter constraint rules, Formula (2) and Formula (3)are the constraints rules of βxandkt.

$$\sum_x \beta x = 1 \tag{2}$$

$$\sum_x kt = 0 \tag{3}$$

According to Equations (2), (3) and the two special definitions of the model, the least squares estimation of the unknown parameters$\alpha x$and$kt$is made and $\alpha x$and$kt$are denoted by symbolsαx^*andkt^*.

$$\alpha_x^* = \frac{\sum_t \ln(m_{x,t})}{n} \tag{4}$$

$$k_t^* = \sum_x k_t[\ln(m_{x,t}) - \alpha_x^*] \tag{5}$$

In them, n is the number of statistical years covered inmx, t, for example, between 1996 and 2016, the value of n is 20 years.

Suppose the estimated value of βx isβx^*. If we want to get regression of $\alpha x^*$and

kt^*, the regression equation is as follows:

$$\ln(m_{x,t}) - \alpha_x^* = \beta_x k_t^* + \varepsilon_{x,t} \tag{6}$$

By Equation (6), the following is available:

$$\beta_x^* = \frac{\sum_t k_t^* [\ln(m_{x,t}) - \alpha_x^*]}{\sum_x (k_t^*)^2} \tag{7}$$

The estimated value of$\varepsilon_{x,t}$ at this time is:

$$\varepsilon_{x,t}^* = \ln(m_{x,t}) - \alpha_x^* - \beta_x^* k_t^* \tag{8}$$

As can be seen from the above steps, the unknown parametersαx, $kt$and$\varepsilon_{x,t}$ are estimated by the least-squares method. Due to the prior estimation ofβx, accurate risk assessment results can be obtained without post-adjustment of the estimation value$\varepsilon x, t^*$of$\varepsilon_{x,t}$. In the big data analysis of health insurance risks, we can get a good fitting effect too. However, the matrix singular value decomposition method used in Lee-Carter model often has the problem of incorrect estimation of the weight of the unknown parameters, and cannot obtain a good fitting result, which causes the result that the subsequent evaluations cannot be accurately carried out.

4.3. Health insurance risk assessment

An insurance company can estimate the threshold of the probability of major illness according to its own profitability. If the result of the assessment of the insurer for a period of time exceeds the threshold, the insurer should be considered of rejection. The risk assessment of health insurance for men and women in our country is made according to $A(0,1,1)$and$A(0,1,2)$, and time parameters of health insurance risk assessment model based on big data analysis in 2016 are estimated as shown in Table 1.

Table 1. The estimates of prevalence of major diseases in 2016

year	k_t^*	
	men	women
2016	-29.276	-35.763

The data in Table 1 and the estimated unknown parameters based onk_t^* were input into the model to assess the risk of $x$age-specific population suffering from major diseases in 2016. As shown in Figure 4, the abscissa is the age range, the ordinate is $m_{x,t}$ evaluation value. At the same time, the real value of the probability of major illness in 2016 is given in Figure 5. By comparing Figure 4 and Figure 5, it can be found that the assessment result of the health insurance risk assessment model based on big data analysis is very close to the true value as a whole. However, there are some subtle differences in the assessment of senior health insurance risk in people over the age of 70, which needs to be amended.

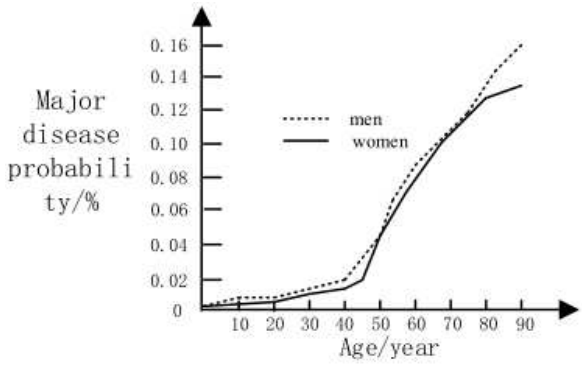

Fig. 4. 2016 major illness probability assessment results

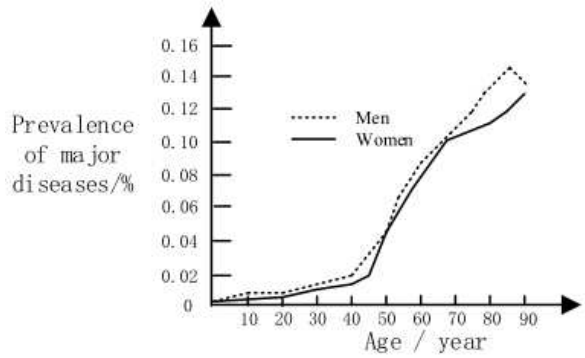

Fig. 5. The true value of major illness in 2016

5. Conclusion

Because health insurance risk is related to social livelihood issues, its complicated big data structure aggravates the necessity of risk assessment. Big data analysis technology can mine the hidden data buried in things and promote the vigorous development of health insurance. The risk of health insurance has the characteristics of generality, similar situation and widespread impact. Under the influence of these three properties, this model is based on the prevalence of the public in recent years in our country, using Lee-Carter improved model for big data analysis of different age groups and assesses health insurance risk by solving model parameters. Simulation experiment organized 50 computers to build a hardware environment and established a big data environment and a high-level virus intrusion environment according to 2016 track2 data collection of international knowledge discovery and data mining competition. Results show that the anti-intrusion capability, false alarm rate and calculation time of this model are more excellent than those of SWM classifier evaluation model and IMU fusion evaluation model, and the overall performance is good.

References

[1] Shen J, He L, Lu H W: *Fractional fuzzy simulation–based health risk assessment for toluene contaminated aquifers.* Human & ecological risk assessment *21* (2015), No. 2, 397–414.

[2] Arunraj N S, Mandal S, Maiti J: *Modeling uncertainty in risk assessment: An integrated approach with fuzzy set theory and monte carlo simulation.* Accident; analysis and prevention *55* (2013), No. 3, 242.

[3] Zhong D, Yang S, Zhang Q: *Filling construction period simulation and risk assessment of high core rock-fill dam based on improved set pair analysis method.* Journal of hydroelectric engineering *34* (2015), No. 3, 137–144.

[4] Zgajnar J: *Simulation model based on iacs data: Alternative approach to analyse sectoral income risk in agriculture.* Tamkang Journal of Science and Engineering *19* (2016), No. 1, 56–64.

[5] Kan G, Yao C, Li Q: *Improving event-based rainfall-runoff simulation using an ensemble artificial neural network based hybrid data-driven model.* Stochastic environmental research and risk assessment *29* (2015), No. 9, 1345–1370.

[6] He D, Hu N, Hu L: *Fault risk assessment of underwater vehicle steering system based on virtual prototyping and monte carlo simulation.* Polish Maritime Research *23* (2016) 97–105.

[7] Hsieh H I, Su M D, Wu Y C: *Water shortage risk assessment using spatiotemporal flow simulation.* Geoscience Letters *3* (2016), No. 3, 1–14.

Received November 16, 2017